Ocean Optics Summer Class

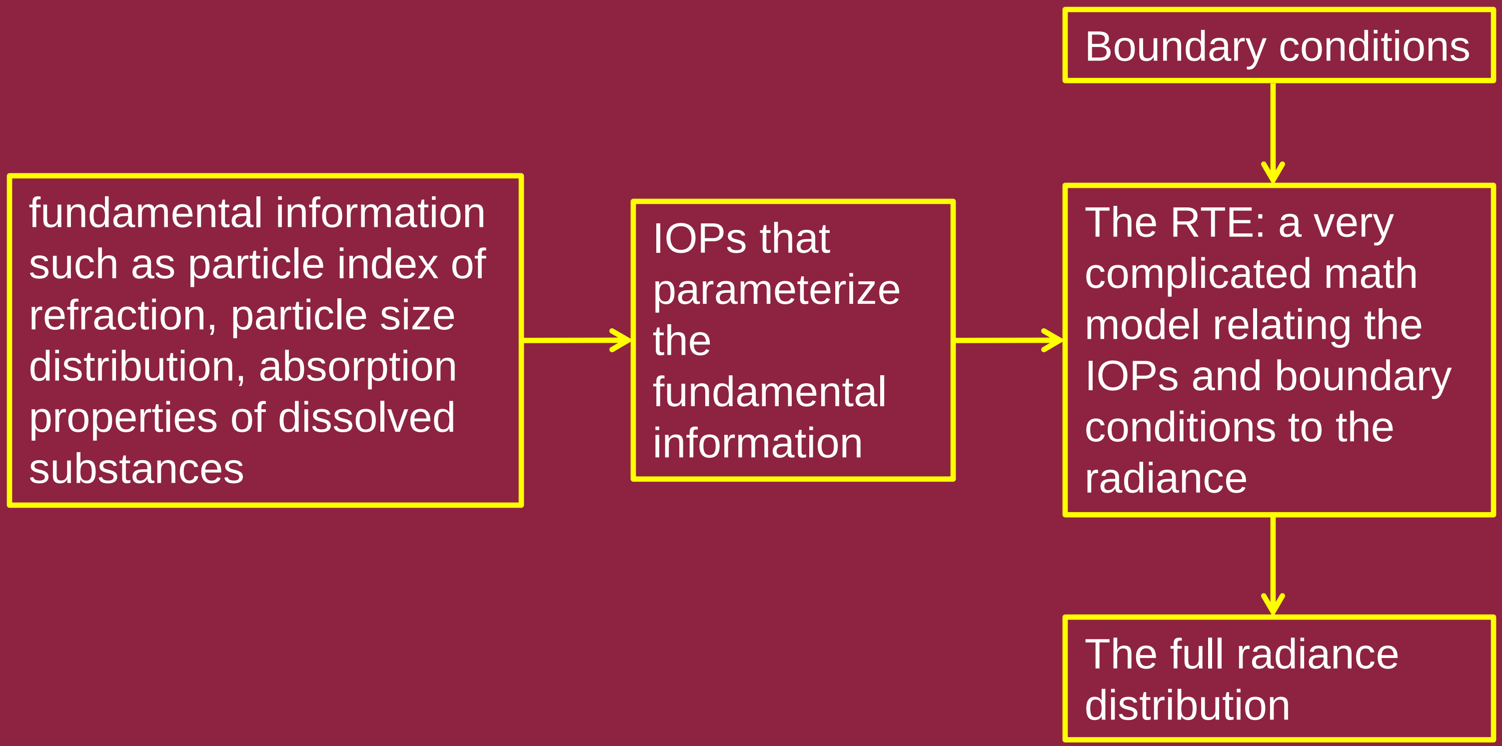
Calibration and Validation for
Ocean Color Remote Sensing

Statistical Methods for Remote Sensing

Curtis Mobley

Delivered at the Darling Marine Center
July 2011

The Radiative Transfer Forward Problem



This is a solved problem: We know how to solve the RTE. All you need is accurate inputs and computer time.

In one viewpoint, the RTE is a fixed, predictive, forward model whose variable input parameters are the IOPs and the boundary conditions, and whose output is the radiance.

Remote Sensing is an Inverse Problem

Inverse problems may have a unique solution *in principle* (e.g., if you have complete and noise-free data), but *they seldom have a unique solution in practice* (e.g., if you have incomplete or noisy data). For example, there may be more than one set of IOPs that give the same R_{rs} within the error of the R_{rs} measurement.

To solve an inverse problem, it is usually necessary to either

(1) add constraints on the solution, to eliminate “wrong” or unphysical mathematical solutions, or

(2) solve for only limited information given the available data (e.g., solve for only a/b_b given R_{rs} or for a and b_b given L_u and E_d)

We always have to worry about non-uniqueness when solving inverse problems, including remote sensing.

Remote-Sensing is an Inverse Problem

imperfect atmospheric correction, unknown boundary condtions

constraints on the allowed solution

A relatively simple math model relating the available light measurements to the IOPs, Chl, bottom depth, etc.

incomplete light measurements: e.g., only R_{rs} at selected wavelengths

an *estimate* of what we want: IOPs, Chl, etc

Statistical Inverse Models

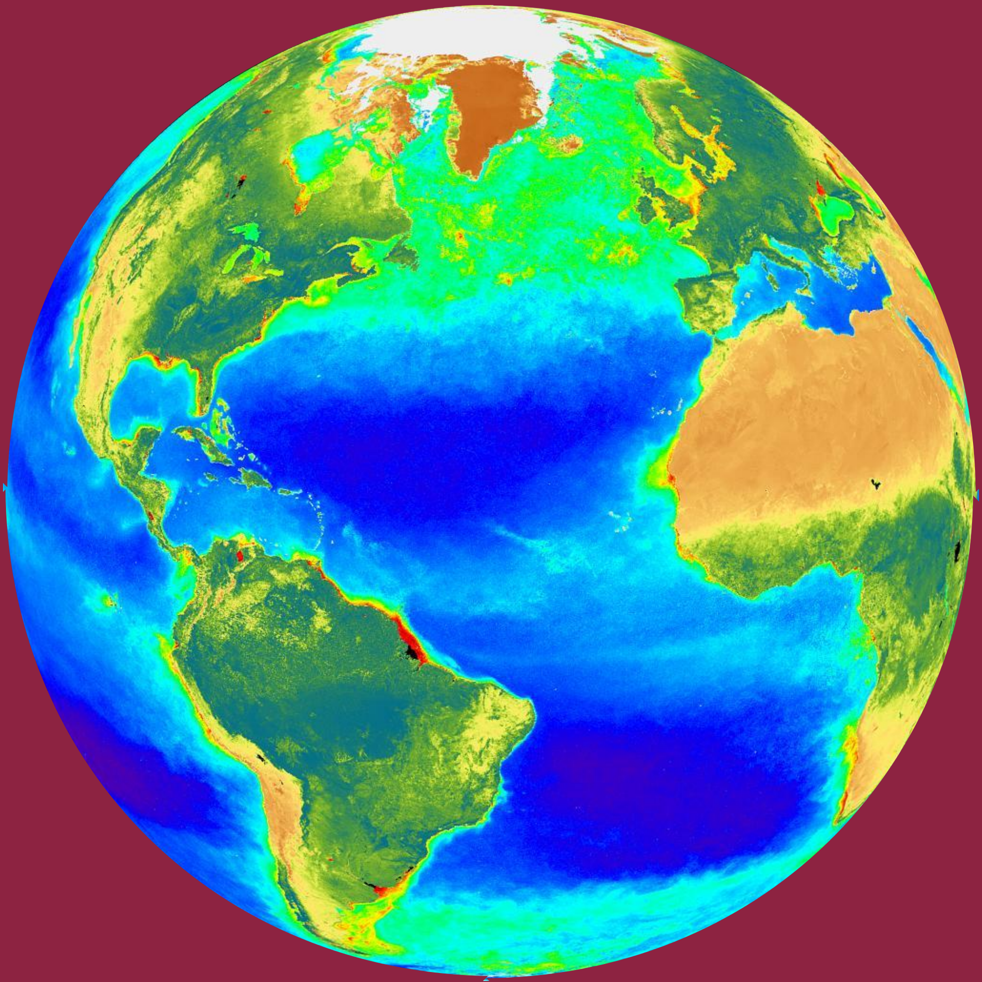
One family of simple math models relating the available measurements to what we want is *statistical* models.

These models are essentially just correlational models obtained from inspection of data sets containing both the inputs inputs (R_{rs}) and outputs (Chl, water depth, etc). The models are not necessarily based on any physical insight as to why the correlation exists.

The general forms of the models contain unknown parameters (proportionality constants, weighting functions, fitting coefficients). The parameter values are determined by *forcing the model to fit data containing both the inputs and outputs*. That is, the parameter values give the statistical best-fit of the model to the data, hence the name "statistical" or "empirical" models.

After the parameters have been determined using known inputs and outputs, the model with the same parameter values can be applied to new input data, to obtain new outputs.

Statistical methods are how ocean color remote sensing got started 40 years ago



Two examples:

- band-ratio algorithms
- neural networks

Where It All Started

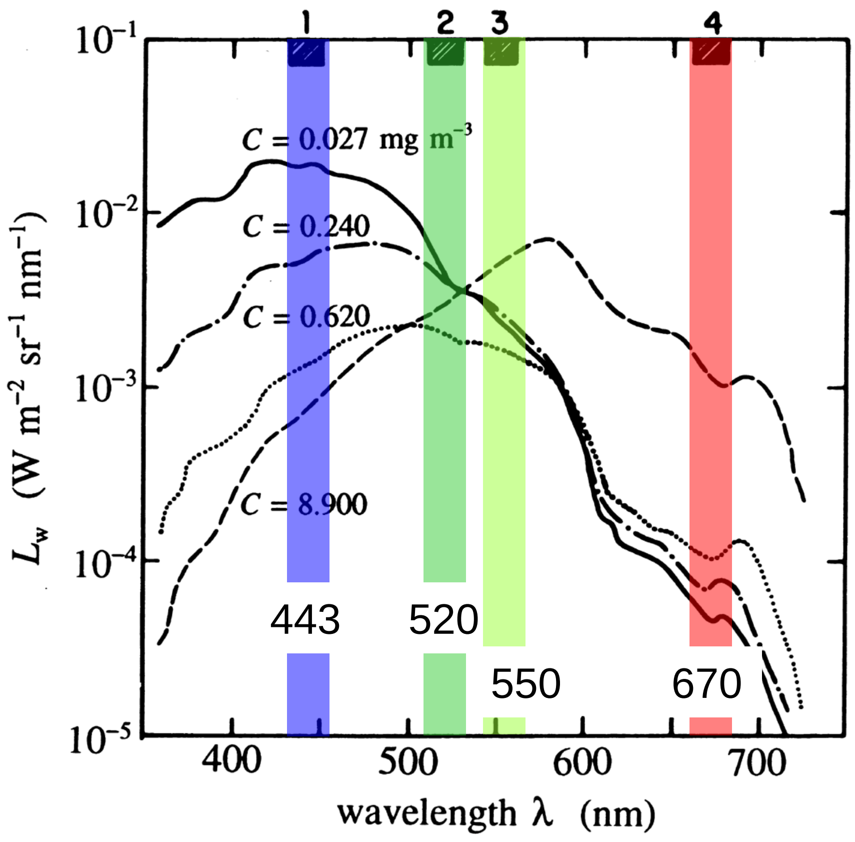
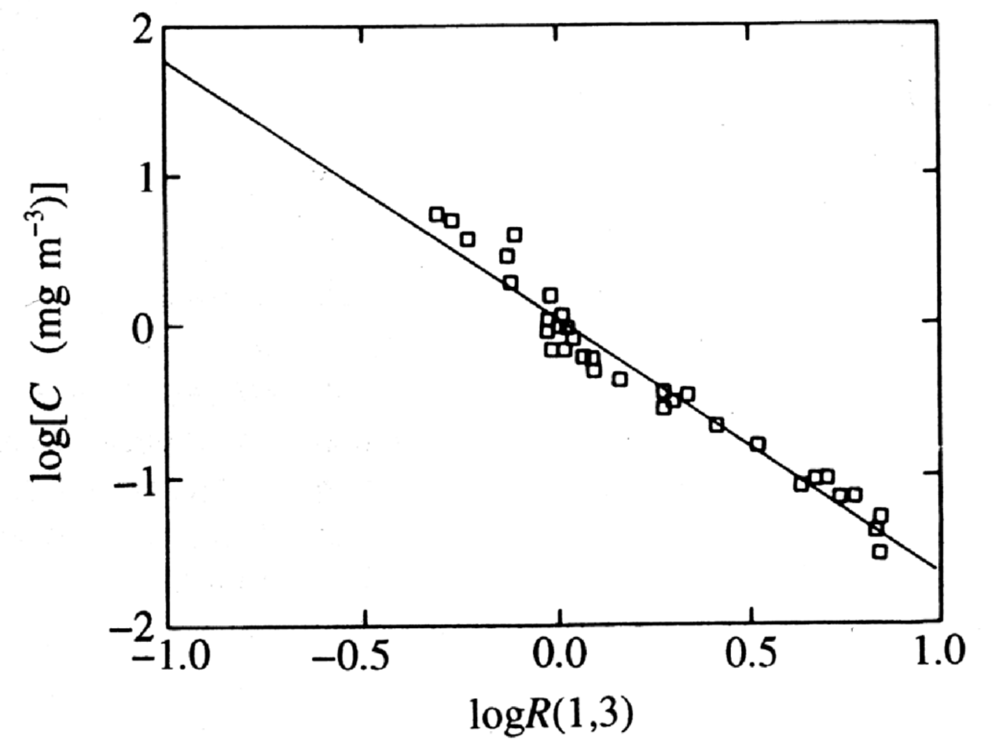


Fig. 10.1. Water-leaving radiances L_w as a function of wavelength for four chlorophyll concentrations C, in case 1 waters. The shaded regions labeled 1-4 indicate the detector bandwidths of the CZCS sensor. [redrawn from Gordon, *et al.*, (1985), by permission]

The seminal idea of ocean color remote sensing: *Chl* concentration and water-leaving radiance are correlated.

$R(1,3) = L_w(\lambda_1{=}443)/L_w(\lambda_3{=}550)$ vs *Chl*



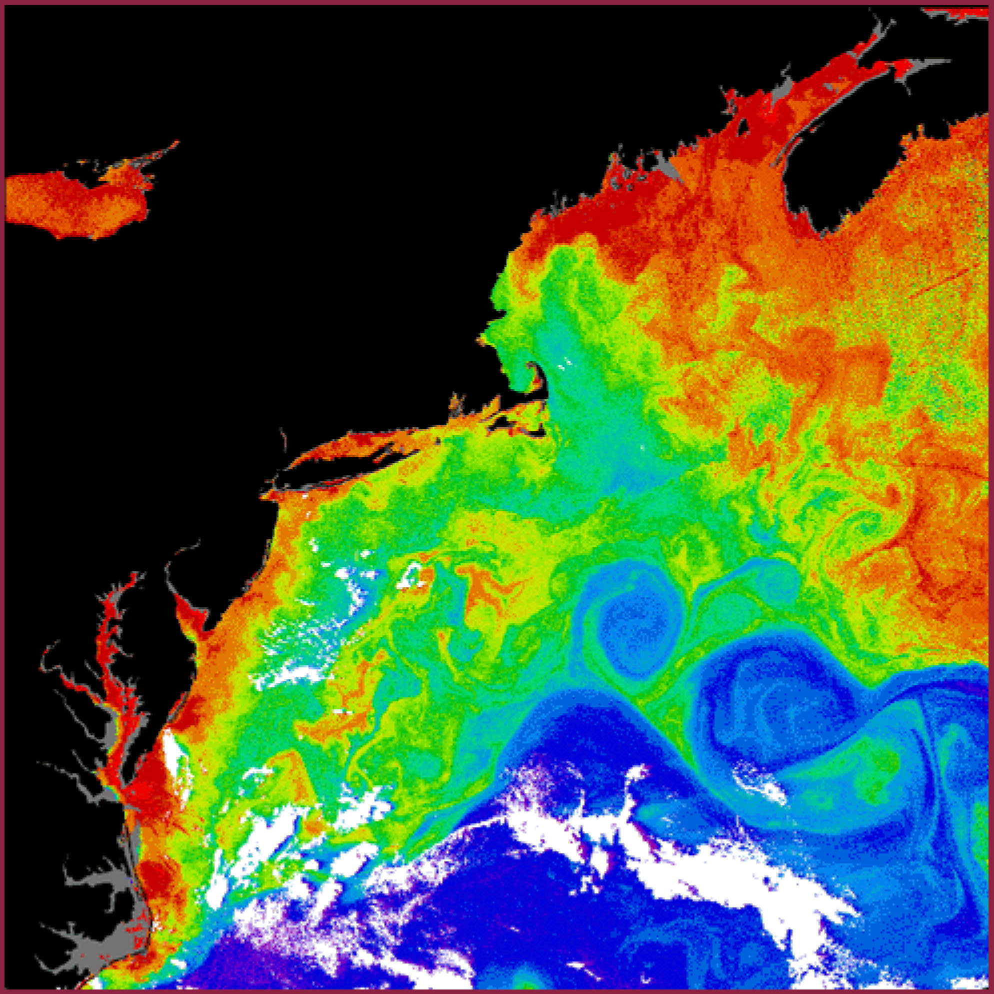
Note: only 33 data points were initially available!

This suggests the band-ratio model:

$\log_{10}(Chl) = C_1 + C_2 \log_{10} [L_w(443)/L_w(550)]$

C_1 and C_2 are the unknown model parameters whose values are determined by a best fit to the data

CZCS Image



Chl = 0.2 in blue to 30 in red

Coastal Zone Color Scanner (CZCS)

1978-1986

4 visible, 2 IR bands

66,000 images

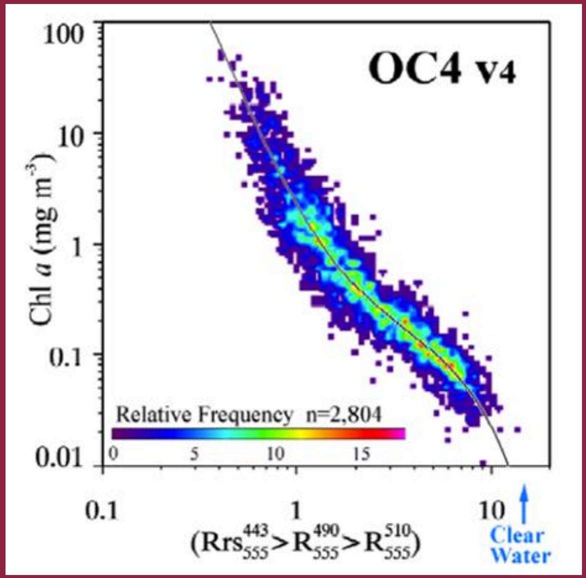
revolutionized oceanography with very simple band ratio algorithms

Examples of Recent Band-Ratio Algorithms

SeaWiFS OC4v4 for Chl:

$X = \log_{10}\{\max[R_{rs}(443)/R_{rs}(555), R_{rs}(490)/R_{rs}(555), R_{rs}(510)/R_{rs}(555)]\}$

$Chl = 10\text{^}(0.366 - 3.067X + 1.930X^2 + 0.649X^3 - 1.532X^4)$



MODIS for $K_d(490)$:

$X = L_w(488)/L_w(551)$

$K_d(490) = 0.016 + 0.156445X\text{^}(-1.5401)$

MODIS for $a_{CDOM}(400)$ and $a_{phy}(675)$:

$r_{15} = \log_{10}[R_{rs}(412)/R_{rs}(551)]$

$r_{25} = \log_{10}[R_{rs}(443)/R_{rs}(551)]$

$r_{35} = \log_{10}[R_{rs}(488)/R_{rs}(551)]$

$a_{CDOM}(400) = 1.5*10\text{^}(-1.147 + 1.963r_{15} - 1.01r_{15}^2 - 0.856r_{25} + 1.02r_{25}^2)$

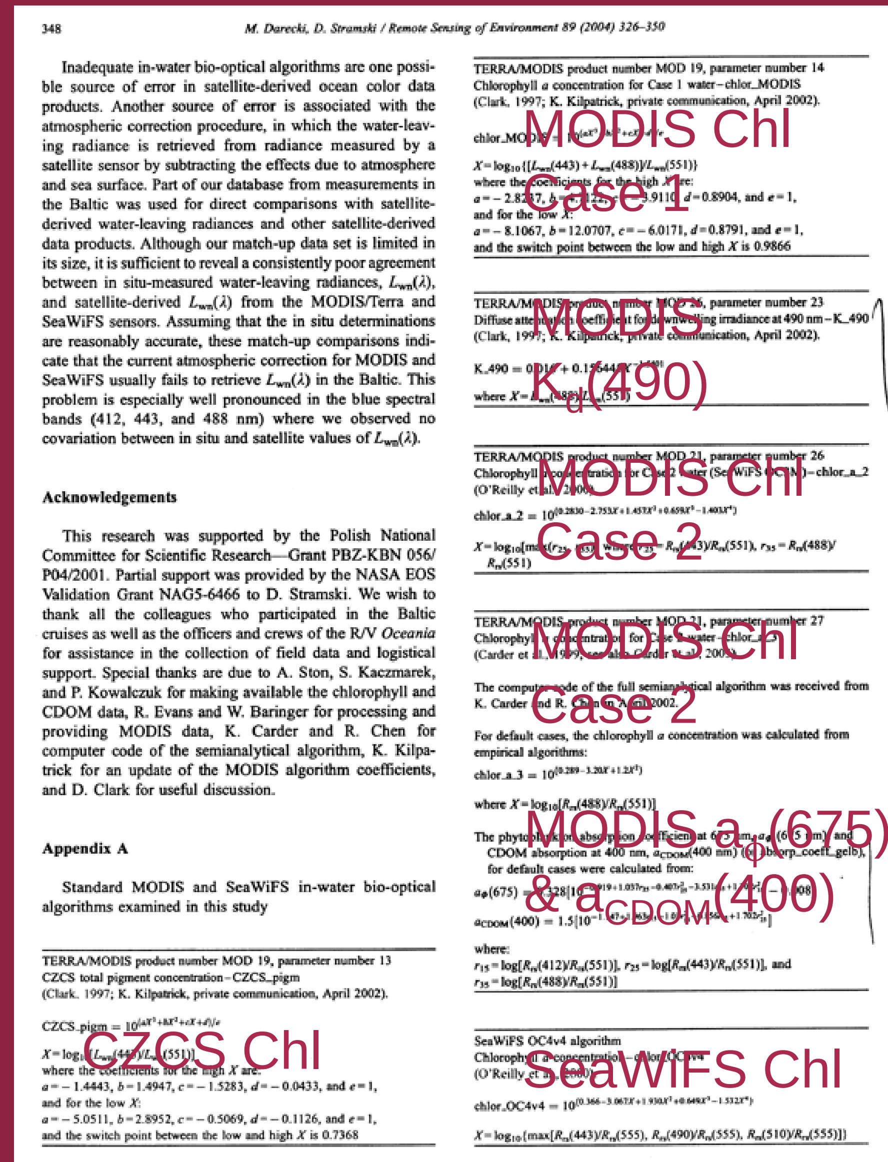
$a_{phy}(675) = 0.328\ [\ 10\text{^}(-0.919 + 1.037r_{25} - 0.407r_{25}^2 -$
$3.531r_{35} + 1.702r_{35}^2 - 0.008)]$

and so on, for dozens more….

A Fun Project

Use HydroLight to generate some R_{rs} spectra for various case 1 and case 2 IOPs. Then run these R_{rs} through various band-ratio algorithms to see how the retrieved values compare with each other and with what went into HydroLight. You can find more on the www.

Darieki and Stramski, *RSE*, 2004



Inadequate in-water bio-optical algorithms are one possible source of error in satellite-derived ocean color data products. Another source of error is associated with the atmospheric correction procedure, in which the water-leaving radiance is retrieved from radiance measured by a satellite sensor by subtracting the effects due to atmosphere and sea surface. Part of our database from measurements in the Baltic was used for direct comparisons with satellite-derived water-leaving radiances and other satellite-derived data products. Although our match-up data set is limited in its size, it is sufficient to reveal a consistently poor agreement between in situ-measured water-leaving radiances, $L_{wn}(\lambda)$, and satellite-derived $L_{wn}(\lambda)$ from the MODIS/Terra and SeaWiFS sensors. Assuming that the in situ determinations are reasonably accurate, these match-up comparisons indicate that the current atmospheric correction for MODIS and SeaWiFS usually fails to retrieve $L_{wn}(\lambda)$ in the Baltic. This problem is especially well pronounced in the blue spectral bands (412, 443, and 488 nm) where we observed no covariation between in situ and satellite values of $L_{wn}(\lambda)$.

Acknowledgements

This research was supported by the Polish National Committee for Scientific Research—Grant PBZ-KBN 056/P04/2001. Partial support was provided by the NASA EOS Validation Grant NAG5-6466 to D. Stramski. We wish to thank all the colleagues who participated in the Baltic cruises as well as the officers and crews of the R/V *Oceania* for assistance in the collection of field data and logistical support. Special thanks are due to A. Ston, S. Kaczmarek, and P. Kowalczuk for making available the chlorophyll and CDOM data, R. Evans and W. Baringer for processing and providing MODIS data, K. Carder and R. Chen for computer code of the semianalytical algorithm, K. Kilpatrick for an update of the MODIS algorithm coefficients, and D. Clark for useful discussion.

Appendix A

Standard MODIS and SeaWiFS in-water bio-optical algorithms examined in this study

TERRA/MODIS product number MOD 19, parameter number 13
CZCS total pigment concentration – CZCS_pigm
(Clark, 1997; K. Kilpatrick, private communication, April 2002).

CZCS_pigm $= 10^{(aX^3+bX^2+cX+d)/e}$

$X = \log_{10}[L_{wn}(443)/L_{wn}(551)]$
where the coefficients for the high X are:
$a = -1.4443$, $b = 1.4947$, $c = -1.5283$, $d = -0.0433$, and $e = 1$,
and for the low X:
$a = -5.0511$, $b = 2.8952$, $c = -0.5069$, $d = -0.1126$, and $e = 1$,
and the switch point between the low and high X is 0.7368

TERRA/MODIS product number MOD 19, parameter number 14
Chlorophyll *a* concentration for Case 1 water – chlor_MODIS
(Clark, 1997; K. Kilpatrick, private communication, April 2002).

chlor_MODIS $= 10^{(aX^3+bX^2+cX+d)/e}$

$X = \log_{10}\{[L_{wn}(443) + L_{wn}(488)]/L_{wn}(551)\}$
where the coefficients for the high X are:
$a = -2.82$[illegible], $b =$ [illegible], $c =$ [illegible]$.9110$, $d = 0.8904$, and $e = 1$,
and for the low X:
$a = -8.1067$, $b = 12.0707$, $c = -6.0171$, $d = 0.8791$, and $e = 1$,
and the switch point between the low and high X is 0.9866

TERRA/MODIS product number MOD 2[illegible], parameter number 23
Diffuse attenuation coefficient for downwelling irradiance at 490 nm – K_490
(Clark, 1997; K. Kilpatrick, private communication, April 2002).

K_490 = 0.016 + 0.15644$X^{-1.[illegible]}$

where $X = L_{wn}(488)/L_{wn}(551)$

TERRA/MODIS product number MOD 21, parameter number 26
Chlorophyll *a* concentration for Case 2 water (SeaWiFS OC[illegible]) – chlor_a_2
(O'Reilly et al., 2000).

chlor_a_2 $= 10^{(0.2830-2.753X+1.457X^2+0.659X^3-1.403X^4)}$

$X = \log_{10}[\max(r_{25}, r_{35})]$, where $r_{25} = R_{rs}(443)/R_{rs}(551)$, $r_{35} = R_{rs}(488)/R_{rs}(551)$

TERRA/MODIS product number MOD 21, parameter number 27
Chlorophyll *a* concentration for Case 2 water – chlor_a_3
(Carder et al., 1999; see also Carder et al., 2003).

The computer code of the full semianalytical algorithm was received from K. Carder and R. Chen in April 2002.

For default cases, the chlorophyll *a* concentration was calculated from empirical algorithms:

chlor_a_3 $= 10^{(0.289-3.20X+1.2X^2)}$

where $X = \log_{10}[R_{rs}(488)/R_{rs}(551)]$

The phytoplankton absorption coefficient at 675 nm, a_ϕ(675 nm), and CDOM absorption at 400 nm, a_{CDOM}(400 nm) (or absorp_coeff_gelb), for default cases were calculated from:

$a_\phi(675) =$ [illegible]

$a_{CDOM}(400) = 1.5[10^{-1.[illegible]}]$

where:
$r_{15} = \log[R_{rs}(412)/R_{rs}(551)]$, $r_{25} = \log[R_{rs}(443)/R_{rs}(551)]$, and
$r_{35} = \log[R_{rs}(488)/R_{rs}(551)]$

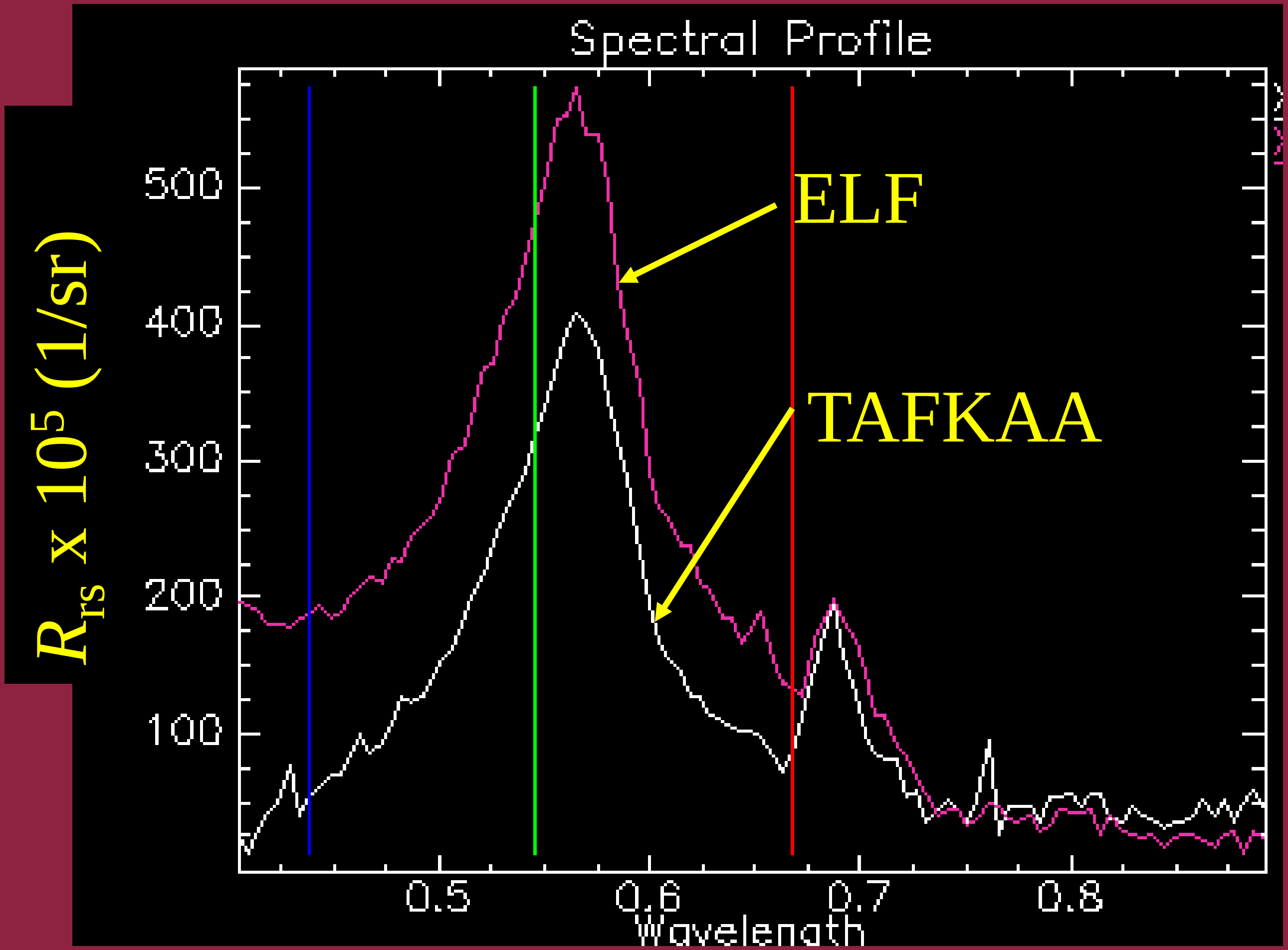
SeaWiFS OC4v4 algorithm
Chlorophyll *a* concentration – chlor_OC4v4
(O'Reilly et al., 2000).

chlor_OC4v4 $= 10^{(0.366-3.067X+1.930X^2+0.649X^3-1.532X^4)}$

$X = \log_{10}\{\max[R_{rs}(443)/R_{rs}(555), R_{rs}(490)/R_{rs}(555), R_{rs}(510)/R_{rs}(555)]\}$

Atmospheric Correction Effects

Band-ratio algorithms can be less sensitive to bad atmospheric correction than some other techniques such as spectrum matching



Nonuniqueness

Band-ratio algorithms are vulnerable to non-uniqueness problems because the R_{rs} ratioing throws out magnitude information that makes spectra unique. Every unique spectrum below has $R_{rs}(490)/R_{rs}(555)$ = 1.71±0.01, which gives *Chl* = 0.59 mg/m^3 by the SeaWiFS OC2 algorithm; all of these spectra had Chl < 0.2 mg/m^3 (the spectra are influenced by bottom reflectance).

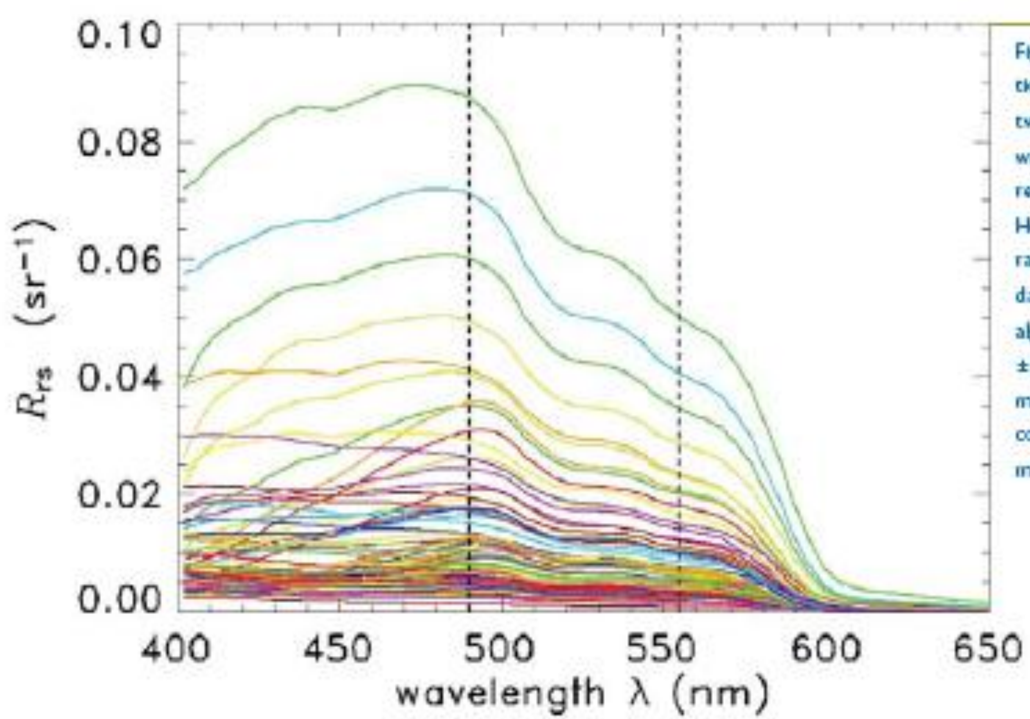
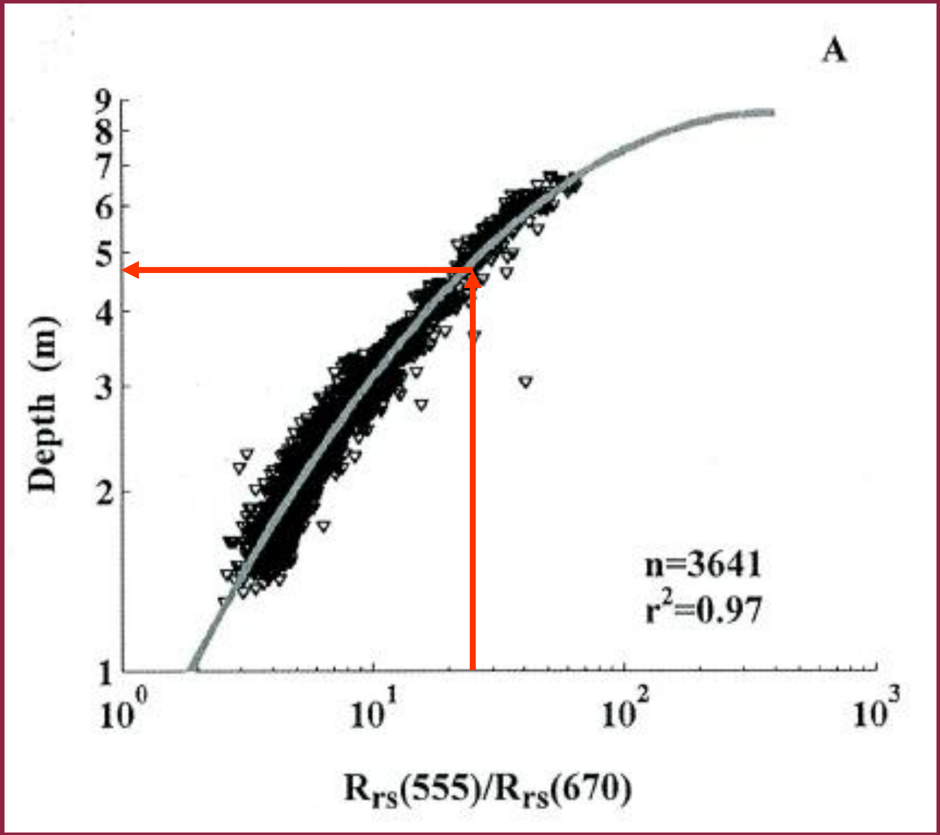


Figure 3. Chlorophyll concentration algorithms designed for multispectral instrumentation may not be useful for shallow, optically clear waters. Shown here are one hundred twenty two Hydrolight-generated remote sensing reflectance (R_{rs}) spectra for Bahamian waters using various combinations of nine different sets of IOPs, 32 different bottom reflectances, and 22 depths between 5.5 and 50 m. These spectra are clearly unique. However, every spectrum has nearly the same remote sensing reflectance wavelength ratio: $R_{rs}(490)/R_{rs}(555) = 1.71 \pm 0.01$ (490 and 555 nm are indicated by the vertical black dashed lines). If this ratio were applied to the commonly used SeaWiFS band-ratio algorithm (OC2; O'Reilly et al., 1998), it would give a chlorophyll concentration of 0.59 ± 0.01 mg Chl m^{-3}. In other words, the same chlorophyll concentration would be determined for all 122 spectra despite the fact that these simulated water bodies have IOPs corresponding to chlorophyll concentrations between 0.0 (pure water) and 0.2 mg Chl m^{-3}. The OC2 algorithm fails here because of bottom effects in optically clear waters.

Nonuniqueness

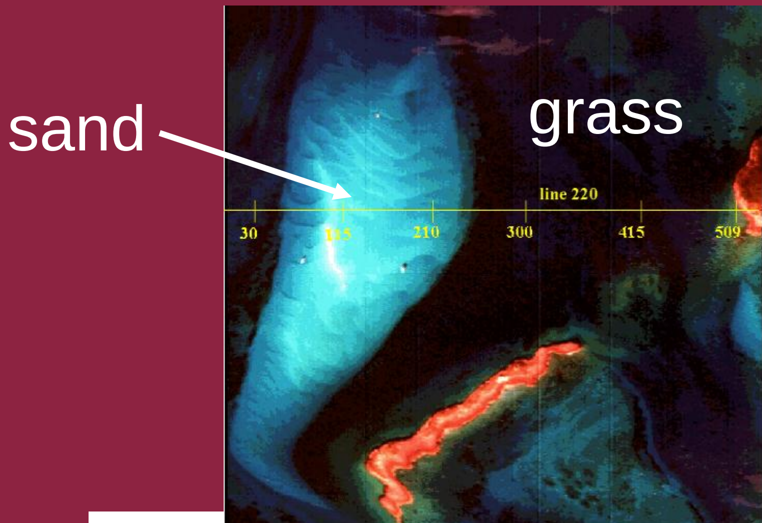
Dierssen et al. (*Limnol. Oceanogr*. 41(1), 444-455, 2003) developed a band-ratio algorithm for bottom depth in clear Bahamas waters:



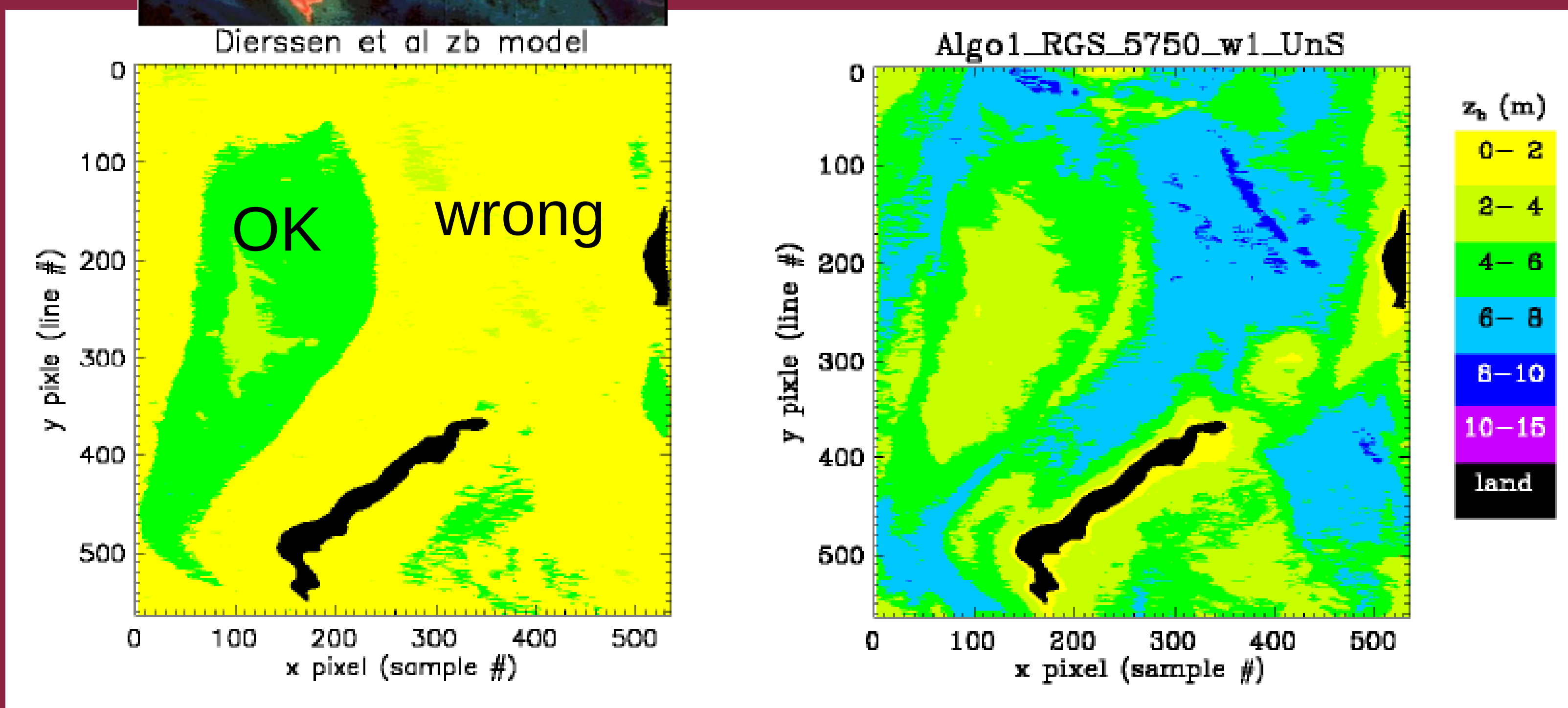
$$x = \log_{10}[R_{rs}(555)/R_{rs}(670)].$$

$$\log_{10}(z_b) = -0.1706\, x^2 + 0.8913\, x - 0.2316.$$

Nonuniqueness

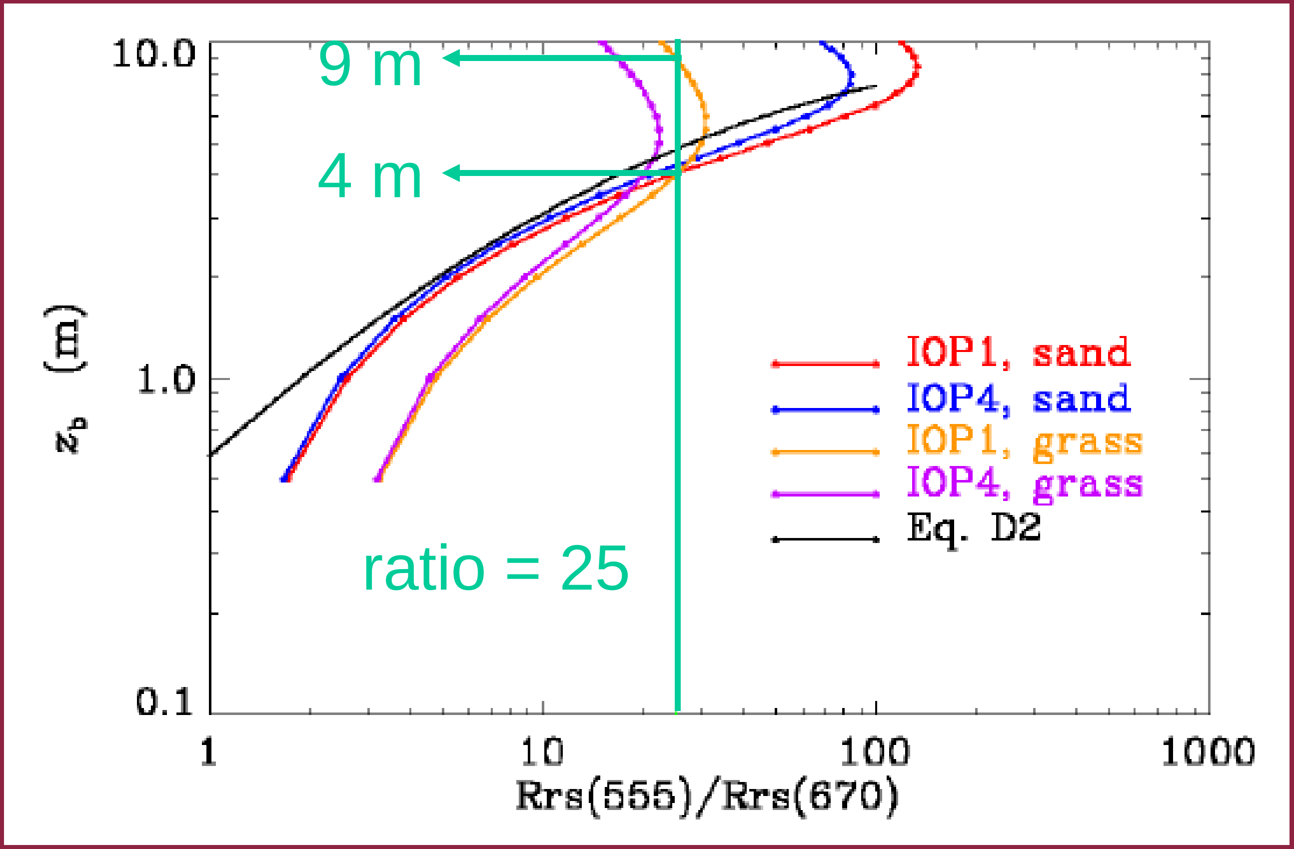


The Dierssen algorithm did OK over shallow sand bottoms, but totally failed over deeper sea grass bottoms. Why?



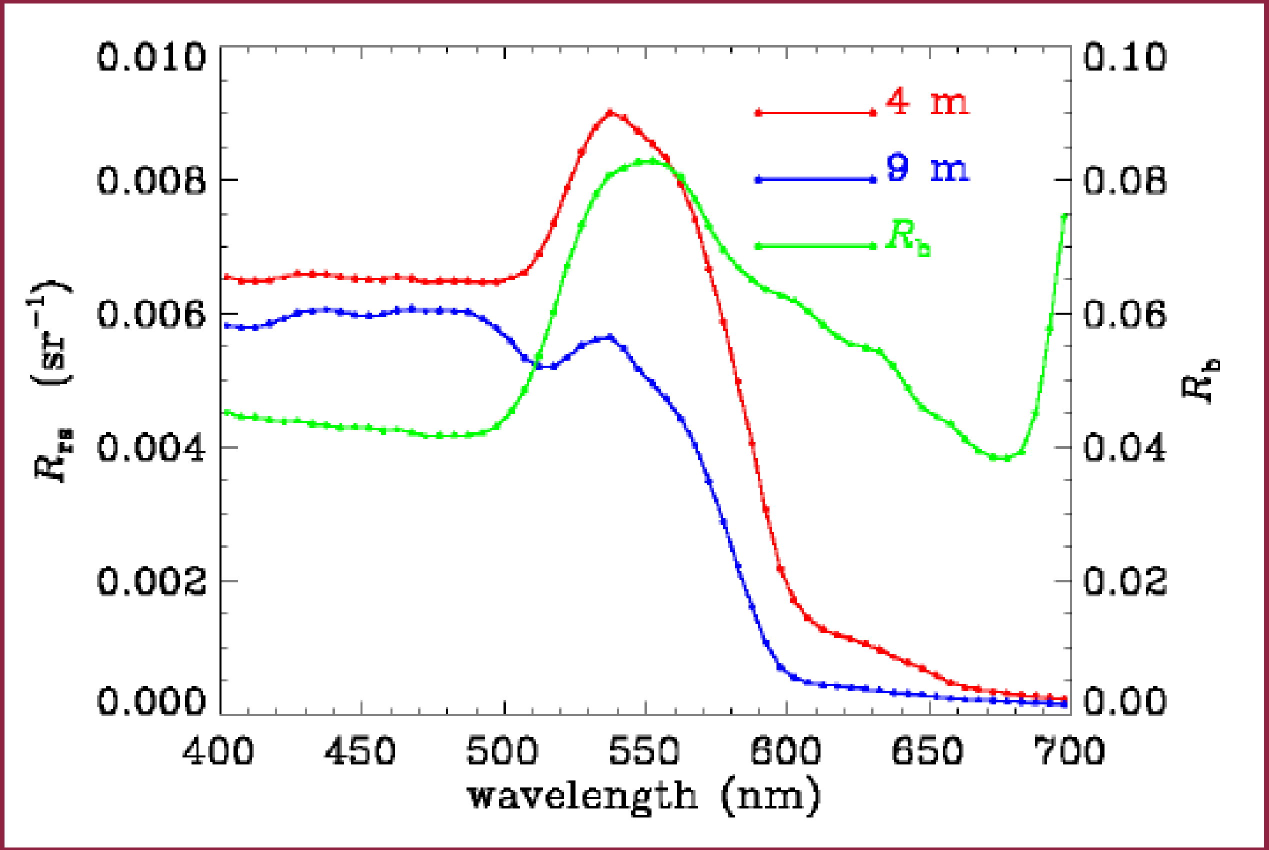
Nonuniqueness

HydroLight simulations of $R_{rs}(555)/R_{rs}(670)$ for two sets of IOPs and two different bottoms (sand and grass), as a function of bottom depth. Nonuniqueness for z_b > 5 m and grass bottom.



Nonuniqueness

The R_{rs} spectra for z_b = 4 and 9 m depth, grass bottom. Both spectra have $R_{rs}(555)/R_{rs}(670)$ = 25 ± 0.1. The Dierssen model gives z_b = 4.8 m.



Model Selection

In some situations, you can figure out (from intuition, theoretical guidance, or data analysis) the general mathematical form of the model that links the input and output (e.g., the polynomial functions that relate the band ratios to Chl). You can then use the available data (e.g., simultaneous measurements of $R_{rs}(\lambda)$ and Chl) to get best-fit coefficients in the model via least-squares fitting.

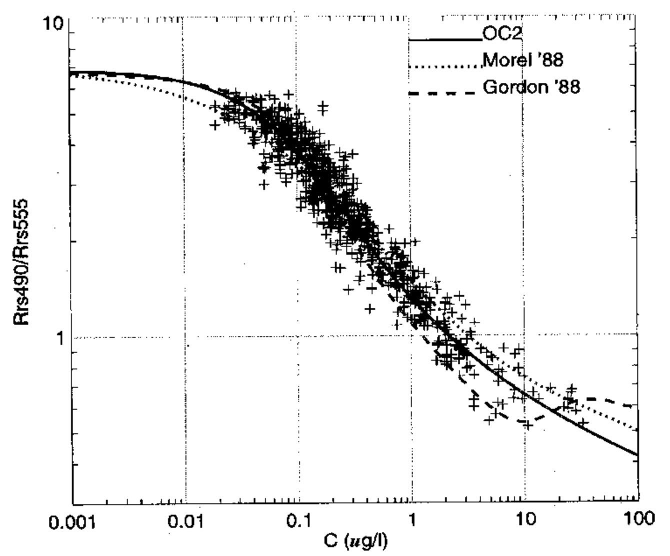


Figure 6. Relationship between chlorophyll and Rrs490/Rrs555 for the ocean chlorophyll 2 empirical algorithm (solid

O'Reilly et al., JGR, 1998

But what if you don't have any idea what the mathematical form of the model is?

Neural Networks

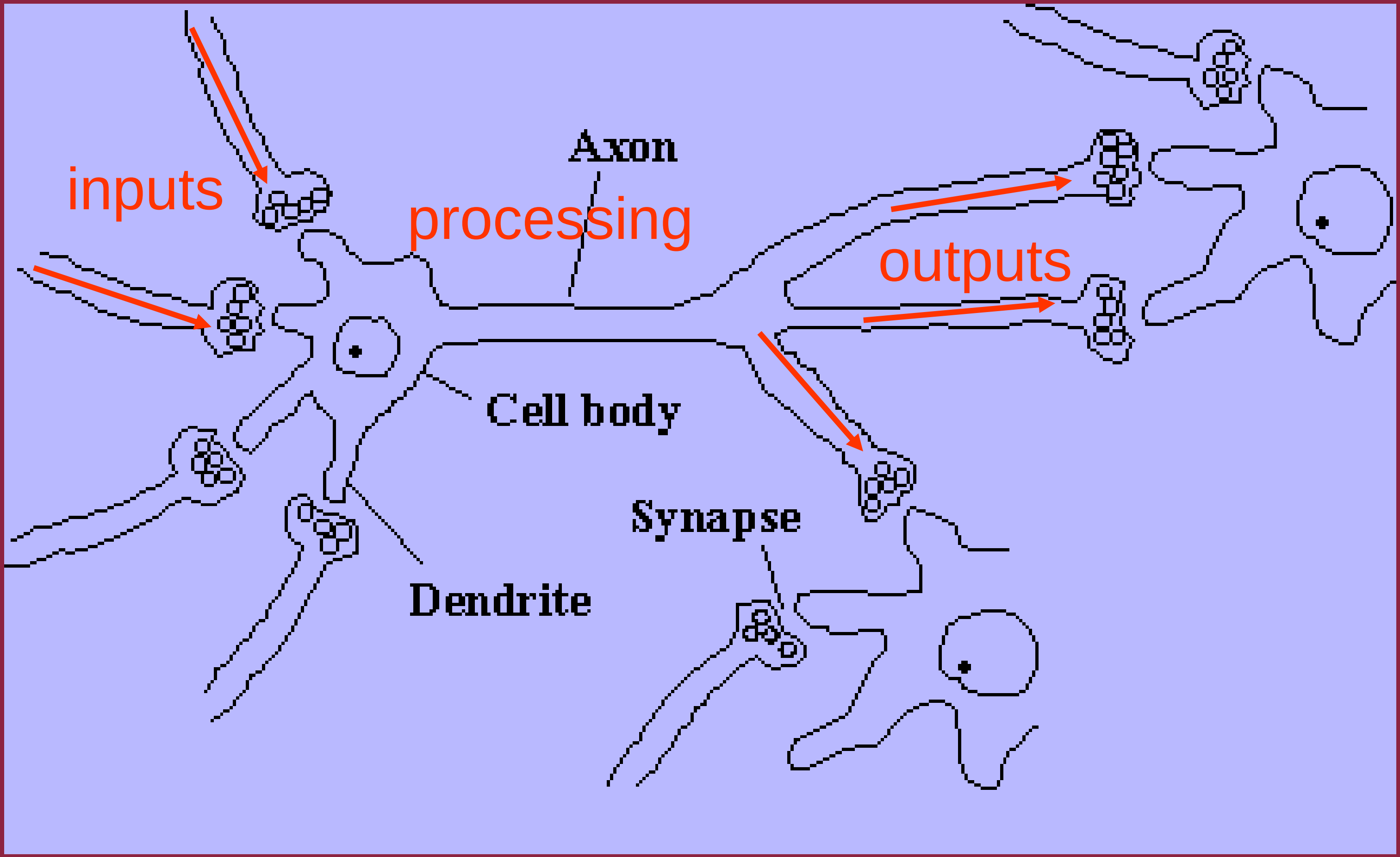
Neural networks are a form of multiprocessor computation, *based on the parallel architecture of animal brains,* with

- simple processing elements
- a high degree of connection between elements
- simple input and output (real numbers)
- adaptive interaction between elements

Neural networks are useful

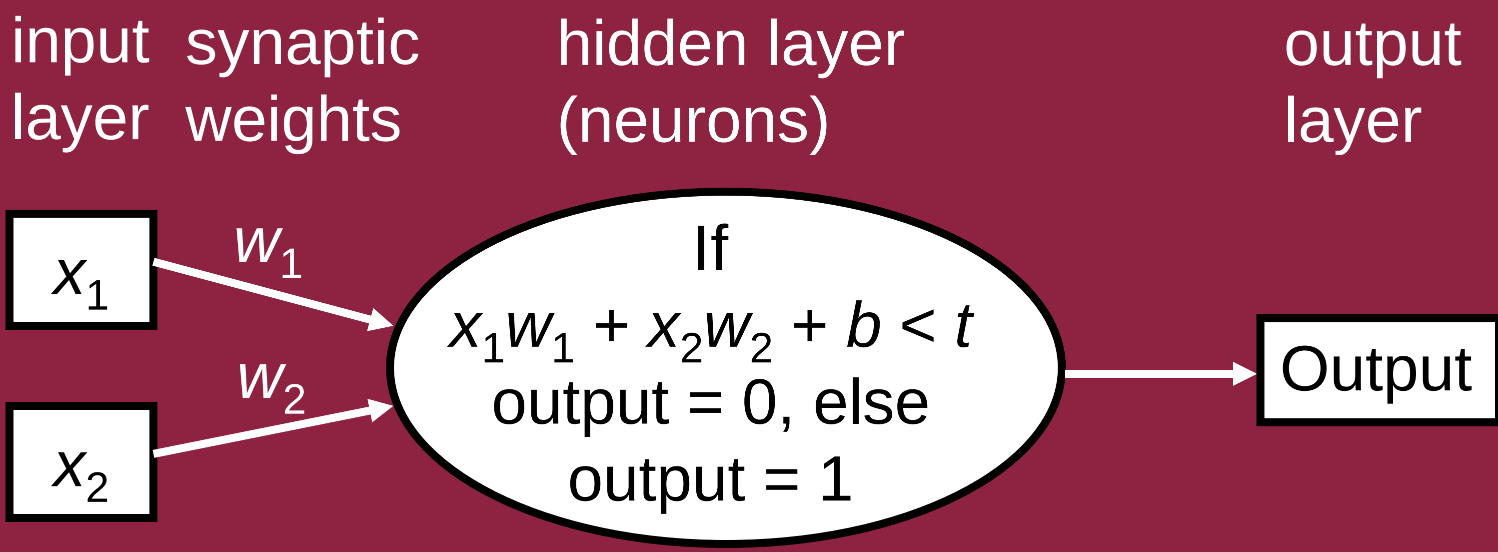
- where we don't know the mathematical form of the model linking the input and output
- where we have lots of examples of the behavior we require (lots of data to "train" the NN)
- where we need to determine the model structure from the existing data

Biological Neural Networks



from www.qub.ac.uk/mgt/intsys/nnbiol.html

A Simple Artificial Neural Network



In the neuron, *b* is the bias, *t* is the threshhold value

The neuron (processor) does two simple things:

(1) it sums the weighted inputs

(2) compares the biased sum to a threshhold value to determine its output

Training the Neural Network (1)

The essence of a neural network is that it can "learn" from available data. This is called *training* the NN. The NN has to *learn* what weighting functions will generate the desired output from the input.

Training can be done by *backpropagation of errors* when known inputs are compared with known outputs. We feed the NN various inputs along with the correct outputs, and let the NN objectively adjust its weights until it can reproduce the desired outputs.

The Java applet at www.qub.ac.uk/mgt/intsys/perceptr.html illustrates how a simple NN is trained by backpropagation.

run the NN applet

Things to Note

The NN was able to use the training data to determine a set of weights so that the given input produced the desired output. After training, we hope (in more complex networks) that new inputs (not in the training data set) will also produce correct outputs.

The “knowledge” or “memory” of a neural network is contained in the weights.

In a more complicated situation, you must balance having enough neurons to capture the science, but not so many that the network learns the noise in the training data.

Training the Neural Network (2)

Another way to train a NN is to view the NN as a complicated mathematical model that connects the inputs and outputs via equations whose coefficients (the weights) are unknown.

Then use a non-linear least squares fitting/search algorithm (e.g., Levenberg-Marquardt) to find the “best fit” set of weights for the given inputs and outputs (the training data).

This makes it clear that NNs are just fancy regression models whose coefficients/weights are determined by fancy curve fitting to the available data (not a criticism!)

An Example NN

From Ressom, H., R. L. Miller, P. Natarajan, and W. H. Slade, 1995. *Computational Intelligence and its Application in Remote Sensing*, in *Remote Sensing of Coastal Aquatic Environments*, R.L. Miller, C.E. Del Castillo, B.A. McKee, Eds.

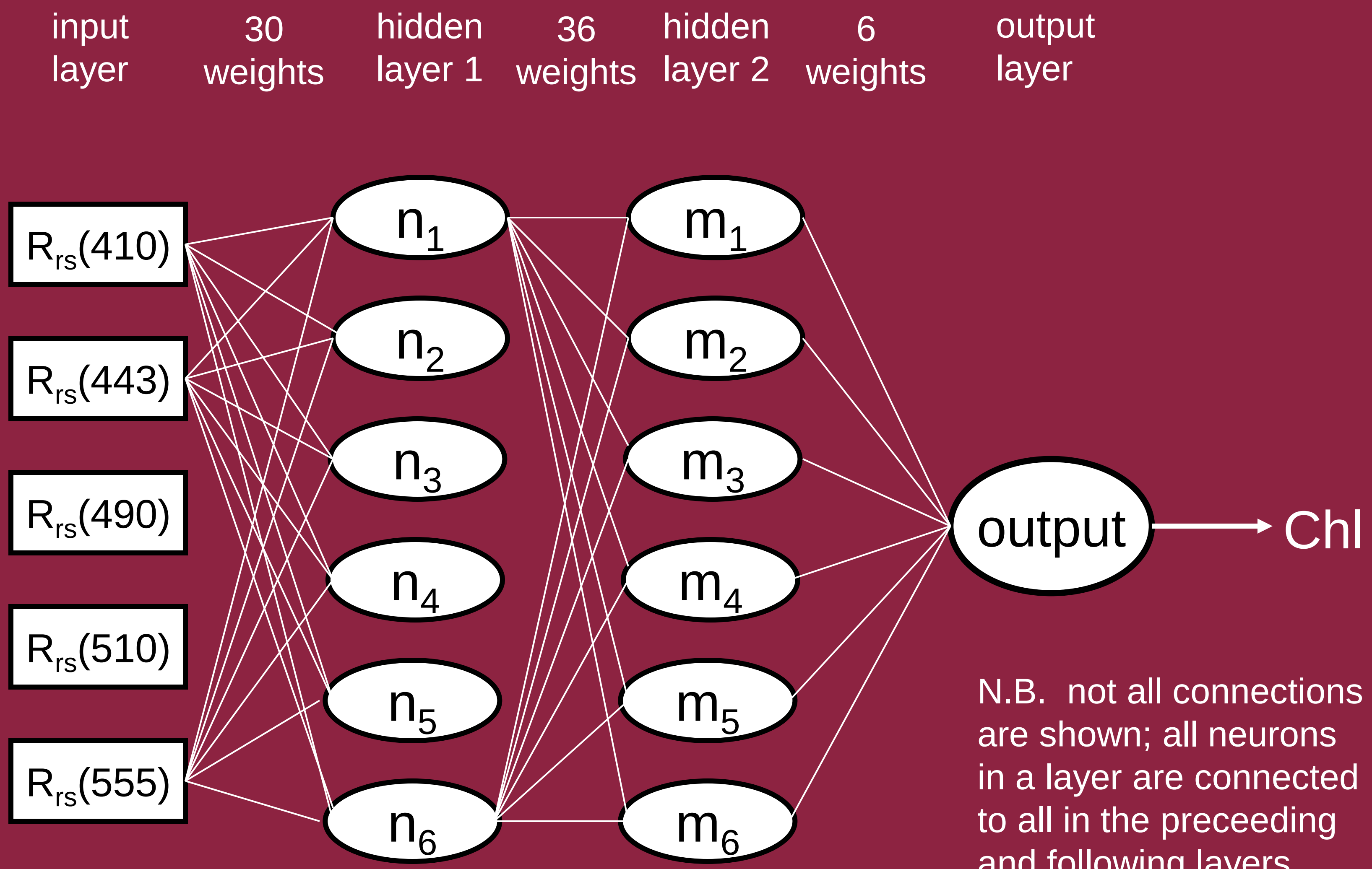
- Assembled 1104 sets of corresponding R_{rs} spectra and *Chl* values from the SeaBAM, SeaBASS, and SIMBIOS databases.

- Construced a NN with 5 inputs (R_{rs} at 5 wavelengths) and two hidden layers of 6 neurons each, and one output (*Chl*).

- Partitioned the 1104 data points into 663 for training, 221 for validation, and 221 for testing the trained NN.

- The NN predictions of *Chl* using the testing data were compared with the corresponding *Chl* predictions made by the SeaWiFS OC4v4 band-ratio algorithm.

The Ressom et al. NN

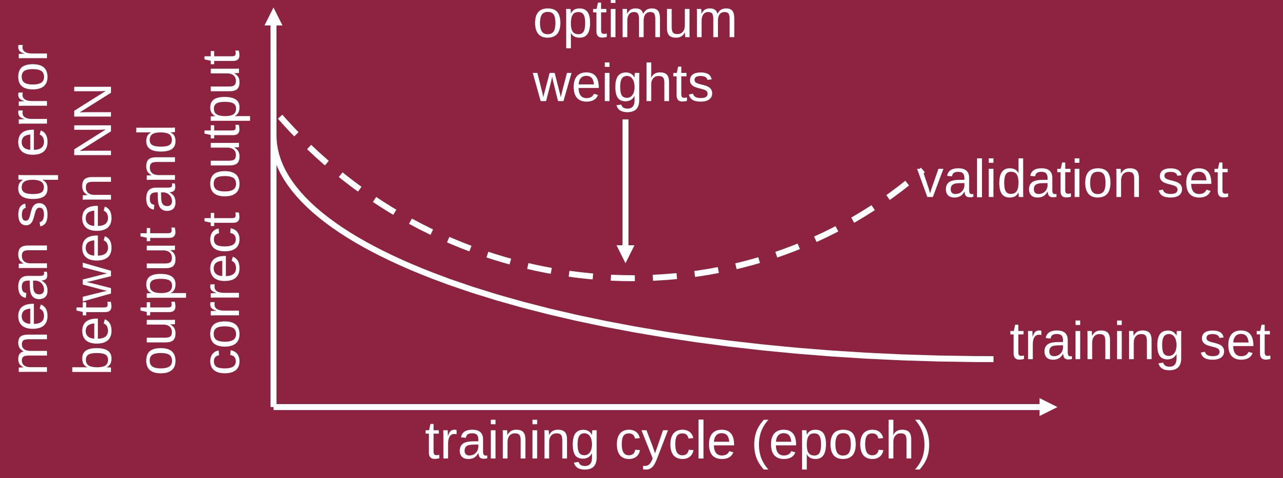


N.B. not all connections are shown; all neurons in a layer are connected to all in the preceeding and following layers

The Ressom et al. NN

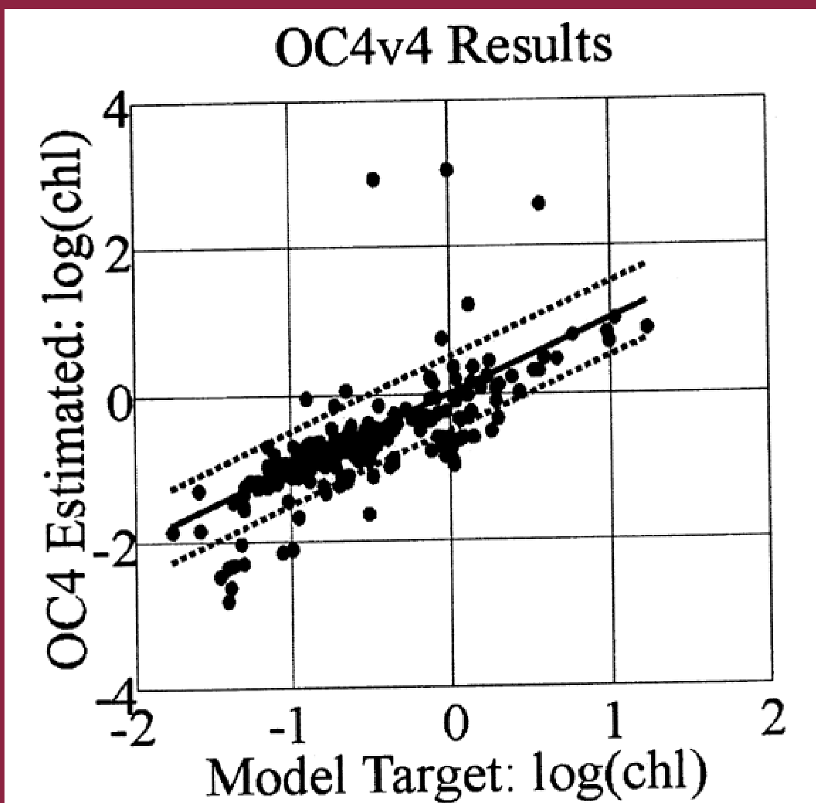
Used two layers of 6 neurons, rather than one layer of 12, (for example), so that neurons can talk to each other (gives greater generality to the NN).

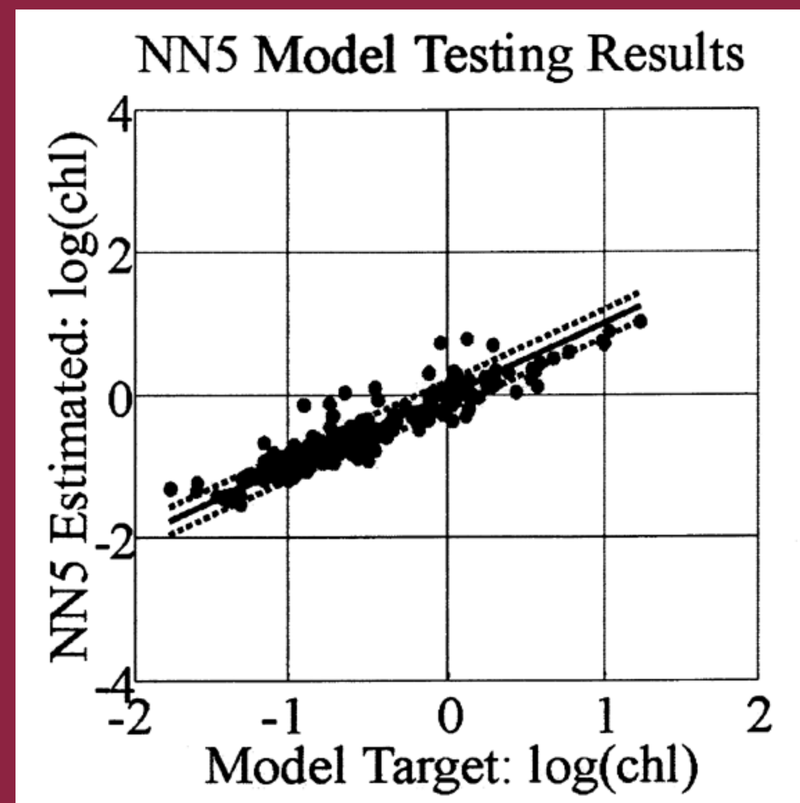
Training uses the training set for weigh adjustments, and the validation set to decide when to stop adjusting the weights.



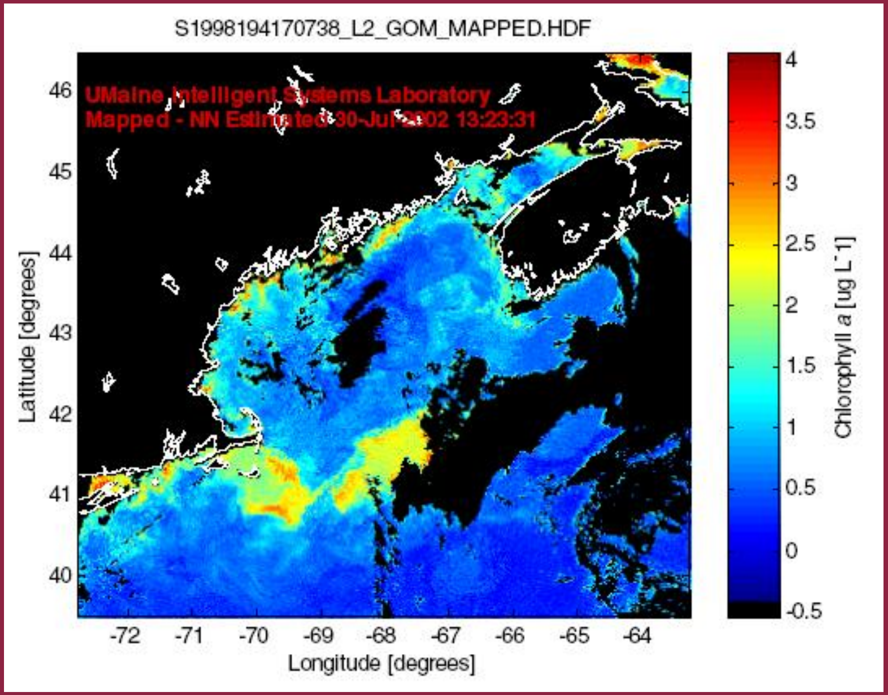
NN vs. OC4v4 Performance

	Training Data (n=662)		Validation Data (n=221)		Testing Data (n=221)	
	r^2	RMS E	r^2	RMS E	r^2	RMS E
OC4	0.651	0.484	0.677	0.450	0.556	0.503
NN5	0.921	0.164	0.837	0.241	0.866	0.199

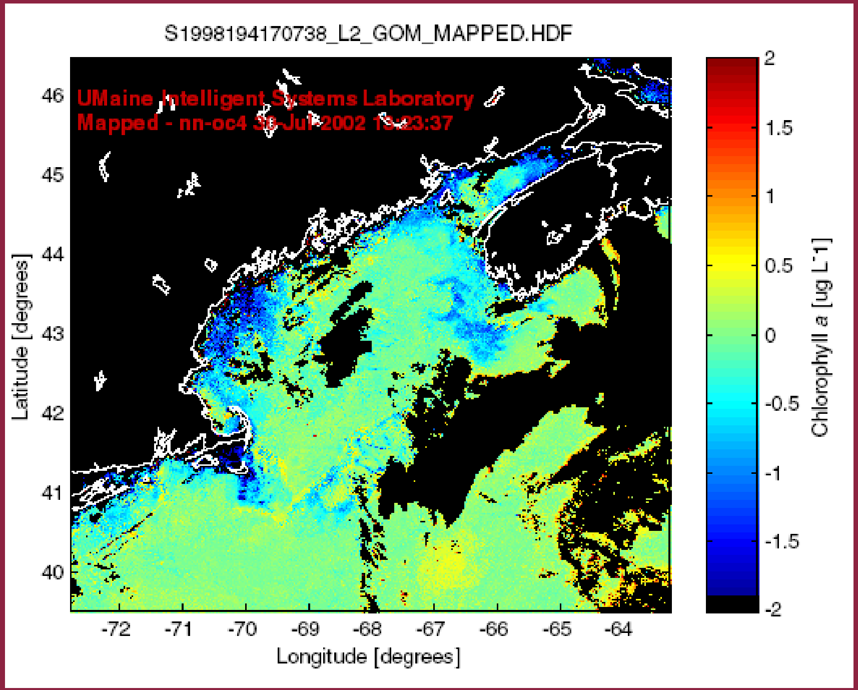




NN vs. OC4v4 Performance



Chl in the Gulf of Maine generated by applying a NN to SeaWiFS data



Difference in the NN and OC4 Chl values (NN-OC4)

from Slade, et al. Ocean Optics XVI

Takehome Messages

Statistical methods for retrieving environmental information from remotely sensed data have been highly successful and are widely used, but...

- An empirical algorithm is only as good as the underlying data used to determine its parameters.

- This often ties the algorithm to a specific time and place. An algorithm tuned with data from the North Atlantic probably won’t work well in Antarctic waters because of differences in the phytoplankton, and an algorithm that works for the Yellow Sea in summer may not work there in winter.

- The statistical nature of the algorithms often obscures the underlying biology or physics.

Takehome Messages

Band-ratio algorithms remain operationally useful, but they have been milked for about all they are worth (IMHO). Note that band ratio algorithms throw away magnitude information in the R_{rs} spectra, and they may not use information at all available wavelengths.

New statistical techniques such as neural networks are proving to be very powerful, as are other techniques such as spectrum matching and semi-analytical techniques.

Limestone (Muav?, early-mid Cambrian, 505-525 Myr old) boulder with fossil algal mats, Grand Canyon, photo by Curt Mobley

The following material gives some standard definitions you need to know

Data Resolution

The quality of remote sensing data is determined by the spatial, spectral, radiometric and temporal resolutions.

- Spatial resolution: The “ground” size of a pixel, typically ~1 m for airborne to ~1000 meters for satellite systems

- Spectral resolution: The number and wavelength width of the different wavelength bands recorded.

- Radiometric resolution: The number of different intensities of light the sensor is able to distinguish. Typically ranges from 8 to 14 bits, corresponding to $2^8 = 256$ to $2^{14} = 16{,}384$ levels or "shades" of color in each band. Useable resolution depends on the instrument noise.

- Temporal resolution: The frequency of flyovers by the sensor. Relevant for time-series studies, or if cloud cover over a given area makes it necessary to repeat the data collection.

Spectral Resolution

Monochromatic:
1 very narrow wavelength band, e.g. at a laser wavelength

Panchromatic:
1 very broad wavelength band, usually over the visible range (e.g., a black and white photograph)

Multispectral:
Several (typically 5-10) wavelength bands, typically 10-20 nm wide

Hyperspectral:
30 or more bands with 10 nm or better resolution
Typically have >100 bands with ~5 nm resolution

Data Processing Levels

- Level 0: Unprocessed instrument data at full resolution

- Level 1a: Unprocessed instrument data at full resolution, but with information such as radiometric and geometric calibration coefficients and georeferencing parameters appended, but not yet applied, to the Level 0 data.

- Level 1b: Level 1a data that have been processed to sensor units (e.g., radiance units), with atmospheric correction applied. Level 0 data are not recoverable from level 1b data. Science starts with Level 1b data.

- Level 2: Derived geophysical variables (e.g., chlorophyll concentration, bottom depth) at the same resolution and location as Level 1 data.

- Level 3: Variables mapped onto uniform space-time grids, usually with missing points interpolated, complete regions mosaiced together from multiple orbits, etc.

Level 4: Model output or results from analyses of lower level data (i.e., variables that were not measured by the instruments but instead are derived from these measurements).